

VOLUME 66
NUMBER 5

WHOLE NO. 337
1952

Psychological Monographs: General and Applied

Combining the *Applied Psychology Monographs* and the *Archives of Psychology*
with the *Psychological Monographs*

HERBERT S. CONRAD, *Editor*

The Validity of Personality-Trait Ratings Based on Projective Techniques

By

HENRY SAMUELS

*Veterans Administration Center
Columbus, Ohio*

Based on a dissertation submitted in partial fulfillment of the requirements for the degree
of Doctor of Philosophy at the University of Michigan

Accepted for publication July 17, 1951

Price \$1.00

Published by

THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.
1515 MASSACHUSETTS AVE., N.W., WASHINGTON 5, D.C.

Psychological Monographs:
General and Applied

COPYRIGHT, 1952, BY THE

AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.

THE VALIDITY OF PERSONALITY-TRAIT RATINGS BASED ON PROJECTIVE TECHNIQUES

<i>Chapter I.</i>	<i>Introduction</i>	1
<i>Chapter II.</i>	<i>Procedures and Methods</i>	2
<i>Chapter III.</i>	<i>Results</i>	7
<i>Chapter IV.</i>	<i>Other Factors Related to Validity</i>	13
<i>Chapter V.</i>	<i>Summary and Conclusions</i>	19
	<i>Bibliography</i>	21

CHAPTER I

INTRODUCTION¹

CONSIDERABLE importance has been attached to projective techniques by clinical psychologists and professional personnel in allied fields. Such devices have been used in psychiatric diagnosis, academic and industrial selection, educational and vocational guidance, research studies of personality development, and in psychotherapy. One of the outstanding problems with respect to such techniques is their validity in their various applications.

¹ I wish to acknowledge my indebtedness to all the people who helped to make this study possible. I am especially grateful to Professor E. Lowell Kelly for his counsel, instruction, and labor, without stint. Such value as may be derived from the study must, for the most part, be attributed to him. I am indebted to Dr. Donald G. Marquis, Dr. George A. Satter, Dr. Max L. Hutt, and Dr. Wm. Clark Trow for their suggestions and criticisms, and to Dr. Donald W. Fiske and Dr. Ernest C. Tupes for their willing help. My wife, Helen, was, as always, inspiring and stimulating, and, in addition, labored long and arduously with the data. Caroline Weichlein and Phyllis E. Moore were most zealous with the secretarial chores.

The opinions expressed herein are the author's and do not necessarily reflect the views of the Veterans Administration.

A review of the literature on projection techniques emphasizes the need for determining the extent to which projective techniques agree with independent criteria in the description of personality. We may have different kinds of validities depending upon the function projective techniques are asked to serve.

The purpose of the present investigation is to continue the study of the validity of projective techniques by considering the following questions:

- (1) How well are projective clinicians who use the same projective method able to *describe* personality?
Are there differences among clinicians in this ability?
- (2) Are there differences in the extent to which personality is correctly described which are related to the kind of projective method used?
- (3) Are there differences among attributes of personality which make for differences in the degree to which they can be correctly described?

CHAPTER II

PROCEDURES AND METHODS

THE present investigation was done as part of the research project on the Selection of Clinical Psychologists sponsored by the Veterans Administration under a contract with the University of Michigan, under the direction of Professor E. Lowell Kelly (3, 4).

A. THE ASSESSMENT PROGRAM

During the Summer of 1947, 140 students who had been selected by universities for first-year positions in the Veterans Administration clinical psychology training program came to Ann Arbor in "classes" of 24 per week to be assessed. During the assessment week a student completed a battery of paper and pencil objective tests, and had administered to him in group form the Thematic Apperception Test (10 cards) and the project's form of a sentence completion test. During the first day or two the student was administered the Rorschach and the Bender-Gestalt individually and at separate sessions. Each student was interviewed twice (6), filled out a 131-item Biographical Inventory, wrote an autobiography, was put through a variety of situations, and filled out a sociometric questionnaire.

The subjects, on whom the ratings investigated in this study were made, consisted of 128 male college graduates who had been accepted by various universities for graduate training in clinical psychology. The twelve women who were part of the total assessed group have been omitted from this study.

Because they are crucial to the present study, a detailed discussion is given to the rating scales, the projective ratings, and the criterion ratings.

B. THE RATING SCALES

The rating scales cover 42 variables divided into three sections which are given the letter designations A, B, and C.¹ Scale A, consisting of 22 variables, was designed to describe more overt or phenotypical dimensions of personality and was adapted from Cattell's (1) factorial studies, with modifications made on the basis of experience with the scale in earlier pilot assessments. In the present study 5 of the 22 Scale A variables are used. These 5 were selected on the basis of a factorial analysis by Fiske (2). Each of these 5 selected Scale A variables has highest factor loadings for Fiske's five factors which are nearly orthogonal to each other. These 5 selected variables are:

- # 4 Depressed-Cheerful
- # 17 Conscientious-Not Conscientious
- # 18 Imaginative-Unimaginative
- # 21 Dependent Minded-Independent Minded
- # 22 Limited overt emotional expression-Marked overt emotional expression

The definitions of Scale A variables were designed to depict these variables as sharply as possible as surface behavior, i.e., how the individual would appear directly to the observer.

The 9 Scale B variables were designed to provide opportunity for judgment of the more covert, genotypical, dynamic, or interpretive aspects of personality. Despite the known intercorrelations between the Scale B variables (from $-.57$ to $.79$) all 9 have been used in this study, since it is in the description of these aspects of personality that projective

¹ Complete definitions of all 42 personality traits may be found in a doctoral dissertation by Samuels (5, Appendix A).

techniques are commonly thought to be most useful. These 9 Scale B variables are:

- #23 Social Adjustment
- #24 Appropriateness of Emotional Expression
- #25 Characteristic Intensity of Inner Emotional Tension
- #26 Sexual Adjustment
- #27 Motivation for Professional Status
- #28 Motivation for Scientific Understanding of People
- #29 Insight into Others
- #30 Insight into Himself
- #31 Quality of Intellectual Accomplishments

Scale C included 11 variables on which staff members were asked to make predictive judgment. Of these, only number 42, "Over-all Suitability for Clinical Psychology," is used here, since this was the variable toward which the entire assessment program was oriented. Throughout the remainder of the discussion Variable 42 is treated as though it were a Scale B variable for simplification in statistical summary. This assumes that the rating on "Over-all Suitability" represents an inference about a covert, although perhaps extremely complex, dimension of personality and is essentially similar to Scale B variables.

All ratings were made on an 8-point scale with a theoretical distribution of ratings of 3, 7, 15, 25, 25, 15, 7, and 3 per cent for points from 1 to 8 respectively. Raters were instructed to use a reference population of first-year clinical psychology graduate students in American universities.

C. THE PROJECTIVE RATINGS

As was stated above, during the early part of the week each student was examined with the Rorschach and the Bender-Gestalt *individually*. Each record was scored, interpreted, summarized, and the applicant rated on the scales described above by the test administra-

tor, without other knowledge about the individual than could be gained from the particular projective technique and the personal contact involved during the testing period. The Thematic Apperception Test and the project form of a sentence-completion test were administered as *group tests*. Each subject was then rated on each variable of Scale A and Scale B by four different staff members, each rater basing his ratings on a different projective technique. A single staff member was concerned with only one projective method per subject and was instructed to keep himself in ignorance of the findings of the other staff members for that subject. After each of the four projective techniques had been independently analyzed and subjects rated on the basis of each, the Rorschach analyst, serving in a new role of "projective integrator," made another set of ratings on each variable of Scale B based on *all* the projective data, with the exception of the actual ratings which had been independently made on the basis of each projective technique.

The number of students seen by a given clinician was not the same for a given technique, nor was the number of students seen by a given clinician the same, necessarily, for two or more techniques. The number of students per rater per technique is shown in Table 1.

A given rater may have analyzed more Rorschachs than another, fewer TAT's than any or all of the others, and some other number of SC's. The differences in *N* between the two Scales, A and B, are attributable to the fact that cases which were integrated by the "projective integrator" were not rated on Scale A by the "projective integrator" nor by the other projective-technique raters. The projective raters used in this study are fewer

TABLE 1
NUMBER OF STUDENTS RATED BY EACH RATER ON FOUR PROJECTIVE TECHNIQUES

Technique	Rater									
	W	O	L	F	T	G	R	Q	M	P
	<i>Scale A</i>									
Rorschach	11	11	12	10	5	9				
TAT	7	8	9	15	3*	14				
SC	4	8	14	4	13	2**				
BG							19	14	15	13
	<i>Scale B</i>									
Rorschach	18	23	22	22	20	20				
TAT	22	22	14	28	4	25				
SC	16	13	37	4	26	6				
BG							32	29	33	33

* With an *N* of three, *r* is either indeterminate or fluctuates widely.

** With an *N* of two, *r* is .00 or 1.00.

in number than the total number of projectivists on the assessment staff. Only six are used, each of whom rated enough students via Rorschach, Thematic Apperception Test, and sentence completion to make comparisons among them possible. Four other projective raters were concerned only with the Bender-Gestalt. The latter four, coded by letters R, Q, M, and P, were graduate students at the University of Michigan, at the third-year level of training in the VA training program for clinical psychologists, and had been trained in the clinical use of the Bender-Gestalt by the same instructor. The other projectivists are coded W, O, L, F, T, and G. They were initially selected on the basis of professional competence, and the choice of another group of projective-analysis raters similarly selected would probably not have resulted in any better interpretation of the projective tests. The sampling of techniques and the sampling of variables are both regarded as adequate. Raters were given full freedom with respect to scoring, interpretation, and report writing. The only formal imposition placed

upon them was the conversion of impressions, deductions, inferences, or conclusions into single numbers on each of the rated variables.

Although students were assigned randomly for projective examination, it was possible that the relatively small number of students seen by a given projectivist using one of the projective devices might have represented a different student population from that seen by another projective rater using the same projective instrument. That is, differences between projectivists' ratings might be found which might be attributable to differences in the students examined and not to differences in the projective techniques. As an additional check on the random assignment of students for projective testing the following test was made. Distributions of the FinP ratings² on students on Variable 42, "Over-all Suitability," were made separately for each of the four projective techniques used in this study. For a given technique

² Final Pooled Ratings. PI will be used as an abbreviation for the Projective Integrator and the Projective Integration.

these ratings for students were distributed for each of the projective raters. In this way it was possible to compare the mean FinP ratings on Variable 42 with respect to the student population seen by each of the projectivists for each of the projective techniques. Epsilon square was computed and no significant differences were found among the mean ratings for students assigned to different projective raters using the same technique. Since the students had been randomly assigned for projective testing and since no differences were found between students for Variable 42, there was no reason to believe that systematic differences would be found between students on other variables, and no other tests were made.

D. THE CRITERION RATINGS

In the seven-day period of each student's assessment, three staff members rated the student at frequent intervals. A given staff member would receive certain data on a student and rate, then receive other data and rate again, having filed away his previous ratings, and continue to make independent ratings based on various kinds of data. The first team conference (Preliminary Pooling Conference) produced a single set of ratings which represented a combination of the judgments of the staff team based on these materials: objective and projective tests, credentials, autobiography and Biographical Inventory, and the information obtained by way of the interview (6). At the final pooling conference, all previous material and all previous ratings were made available to the staff team. Also available at this time were the student's self-ratings, ratings of each student by the other three members of the student team, the separate ratings

based on projective methods, and the ratings made by a team of staff members who had observed the student in situation tests with no other knowledge about him. A separate final pooled rating was made on each variable of Scale A and of Scale B. These ratings are regarded as the best available measures of each of the 42 personality traits for each subject assessed, and are the criterion measures of this study. They are designated FinP.

The FinP ratings are the most comprehensive and inclusive ratings to come out of the entire week's assessment. These ratings were made by different combinations of three staff members. The staff were initially selected on the basis of professional competence. They had had opportunity to study the student in a wide variety of ways, and had had the benefit of judgments of many other skilled clinicians at various points. Although they are admittedly fallible, there can be little doubt that these FinP ratings are as valid criterion measures of these variables as are obtainable at the present time from skilled clinicians using present techniques in an assessment situation.

The FinP ratings are composite criteria consisting of the pooled judgments of three staff members who had available to them the results of the tests and procedures indicated above. A separate FinP rating was made for each variable of Scale A and Scale B. Since staff members were free to arrive at their final judgments as they chose, it would be difficult to say whether FinP ratings emphasized abilities, capacities, past achievement, or personality characteristics. Of interest in this connection is "... the fact that assessment staff members tended to be uniformly of the opinion that the interview contributed most to their understanding

of the case,' followed by either the projective tests or autobiography" (4, p. 404).

Two problems must be considered with respect to the relationship between the ratings based on projective techniques and the FinP ratings. First, how much influence did any or all of the projective techniques have on the criterion measures? Second, are there differences in the criterion ratings when the Projective Integrator is and is not a member of the criterion team?

With respect to the first problem of the influence of projective techniques on criterion measures, a quantitative answer cannot be given. The projective technique protocols and the ratings based on them were available to the final pooling team. This constitutes a theoretical defect in the design of this study which was a sacrifice which had to be made in consideration of other aspects of the over-all project design. It would have been preferable, of course, if the ratings based on each projective technique could have been checked against a criterion based on all available data except that technique. This deficiency with respect to experimental independence would have the effect of spuriously raising the correlations between the ratings based on projective techniques and the criterion. The possible effect of dominance in criterion ratings by any one criterion team member tended to be cancelled by the rotation of staff members from week to week in such a way that for each of the student classes the criterion teams were differently constituted. This does not eliminate the possibility of what might be called group bias with respect to projective methods. For example,

many of the staff members, not themselves projectivists, spoke very respectfully of the value of the Rorschach and may have given careful attention to the Rorschach ratings in arriving at the criterion ratings. The possibility exists that another staff, or the same staff assessing at another time, might have another attitude toward projective methods and arrive at a different set of criterion ratings for the same subjects. Since there is small likelihood of another assessment designed to measure the reliability of criterion ratings by the test-retest technique, these obtained criterion measures must be regarded as the only ones available, and therefore, in a practical sense, as universe (not sample) measures.

The problem of the effect of the Projective Integrator as a criterion team member on the criterion team ratings can be evaluated. The Projective Integrator was a member of the criterion team for half the number of poolings. Since the Projective Integrator was also the Rorschach analyst, it might be hypothesized that if he exerted a constant influence on the criterion ratings in the direction of his Rorschach ratings, then the correlations between ratings based on Rorschach and criterion ratings would be higher when the Projector Integrator was a member of the criterion team than they would when he was not. A comparison of correlations between Rorschach ratings and FinP ratings when the Projective Integrator was and was not a member of the criterion team found none of the differences between correlation coefficients under these conditions to be significant at the .05 level, nor was the direction of difference consistent.

CHAPTER III

RESULTS

A. THE PROJECTIVE TECHNIQUES

1. *Validities of ratings based on projective techniques*

TO DETERMINE the extent to which ratings based on projective techniques correlated with FinP ratings, the ratings made by the six raters (four in the case of the Bender-Gestalt) were entered into a single scatter plot for each technique and variable. It was found that something other than "chance" operated to generate statistically significant correlations, since, for Scale A variables, only 1 correlation in 20 was expected to be significant at the .05 level by chance and we find 9 in 20 significant at this level. With respect to Scale B, where 40 correlations were computed, we would expect 2 to reach the .05 level and we find 30 of the 40 correlations are significant at this level. In addition it was noted that only 2 of the 60 coefficients were negative. These two findings, the relatively large number of statistically significant correlations and the high proportion of positive correlations, suggest that the ratings based on projective techniques have some validity. However, these results are, to some extent, spurious since, in addition to the fact that the criterion measures by the six raters were not completely independent, the inter-

correlations among Scale B variables were more often positive than negative, and fairly high.

The median correlation coefficients between ratings based on each of four projective techniques and FinP ratings are presented in Table 2. It was noted that correcting the technique-FinP r 's for attenuation in the FinP ratings raised the median r only slightly, the greatest increase being from .31 to .35 in the case of the SC on Scale A. That is to say, that even had the FinP ratings been perfectly reliable, the ratings based on projective techniques would have correlated only slightly higher with the FinP ratings.

2. *Differences in the validities of ratings among the projective techniques.*

In order to test the hypothesis that there are no differences in the validities of ratings made on the basis of the four techniques used in this study the following analysis was made. The validity coefficients for each of the techniques for each of the 15 variables were transformed into Fisher's z function and analysis of variance was done for each of the two sets of variables. There was no rater variance here since raters had been combined for each technique. The results of the analysis of variance for the transformed validity coefficients for tech-

TABLE 2
MEDIAN OF CORRELATIONS BETWEEN RATINGS BASED ON EACH OF FOUR PROJECTIVE TECHNIQUES AND FINAL POOLED RATINGS FOR SCALE A AND SCALE B VARIABLES

Scale	Technique							
	Ror.	<i>n</i>	TAT	<i>n</i>	SC	<i>n</i>	BG	<i>n</i>
Scale A	.20	58	.32	56	.31	45	.21	61
Scale B	.29	125	.26	115	.28	102	.14	127
Both Scales	.27		.27		.28		.19	

niques for Scale A showed no significant differences in validity at the .05 level for the techniques or for the variables. The analysis of variance of the Scale B variables showed no significant differences between variables but a difference between techniques significant at the .01 level. Further analysis showed that this difference was attributable to the ratings based on the Bender-Gestalt; this technique differed significantly from each of the other three techniques which did not differ significantly among themselves. Since the Bender-Gestalt raters were four different people from those whose ratings were based on the other three techniques, conclusions from these findings must be made with great care.

3. *Comparison of technique validities on Scales A and B*

When the median validity coefficients, which are given in Table 2, are examined, one is struck by the fact that with the exception of the Rorschach the coefficients are somewhat higher for Scale A, which is not a result that might have been readily hypothesized beforehand. It may also be noted that while many of the median values are statistically significant (.05 level) none of them may reasonably be regarded as predictively useful. Thus, for the techniques studied (with the exception of the Rorschach), we see that projective methods do not lead to more valid inferences about the more presumably covert aspects of personality, as is frequently argued, than they do of the more obvious aspects of personality. On the other hand, and particularly with reference to the TAT and the SC where the student was not seen by the rater (Ror and B-G were administered by the rater) and behavioral manifestations of the student were in-

ferred from the protocol only, we find that ratings are as valid, by and large, for overt variables as for covert variables.

The possibility that these findings may have been related to the unreliability of the criterion ratings has been shown to be dismissible. Another criticism of these findings may ask if it is fair to require clinicians (in this instance projective clinicians) to express their impressions and judgments in the form of ratings. The core of the argument is that clinicians are accustomed to describing their findings in the form of language (interpretive findings beyond the scores, ratios, percentages, etc. required by the various test instruments) and not in the form of ratings. While this problem must, perhaps necessarily, remain largely a matter for individual judgment, there is considerable merit in the argument that numbers may be more precise than words, and are definitely far more useful in the verification of hypotheses. This position does not deny the possibility that ratings may not accurately represent the judgments about people which clinicians can make from personality test data, nor does it deny the possibility that such judgments may be more accurately communicated in some form other than ratings.

That there are not more differences in validities between the two scales for the projective techniques is a curious finding and difficult to understand. It is not unreasonable to believe that if one has understanding of the *dynamics* of behavior one can infer one step more to the overt behavior of a given individual, but in making the additional inferential step there is additional opportunity for error. It would not have been surprising to find that ratings of overt personality characteristics had less relationship to the

criterion ratings than ratings of the covert traits. That they do not—and since in both cases the relationships with the criterion ratings are low—suggests further study to determine, for example, how much correlation there is between the criterion measures and ratings based on observations of the subject in the projective testing situation, and how much with ratings based on projective data without direct observation of the subject.

4. Correlations between projective technique ratings

Up to this point the four projective techniques have been considered as generically related. They have been called by a class name and regarded as methods of evaluating personality characteristics.

We have found that ratings based on each of the four techniques show but low correlation with the criterion measures. We shall turn to the question, to what degree are personality ratings based on each of the separate techniques in agreement? This was investigated by intercorrelating the ratings, variable by variable, for each of the four techniques. The results are shown in Table 3. While there is a tendency toward positive relationships (101 out of 120) among the independent ratings (no two ratings of a single subject on a given trait were made by the same projective technique rater), there is very little basis for a feeling of confidence that these instruments serve similar functions, i.e., that these instrument-clinician combinations are measur-

TABLE 3
CORRELATIONS BETWEEN RATINGS BASED ON EACH OF THE PROJECTIVE TECHNIQUES*

Variable	Scale A					
	Ror/TAT	Ror/SC	Ror/BG	TAT/SC	TAT/BG	SC/BG
4	.02	.08	.08	-.05	.11	.14
17	.11	-.01	.11	.09	.13	-.01
18	.11	.21	.10	.23	.06	.09
21	-.20	.22	.06	.05	.10	-.02
22	-.11	-.01	.27	.46	.03	.04
Variable	Scale B					
	Ror/TAT	Ror/SC	Ror/BG	TAT/SC	TAT/BG	SC/BG
23	.12	.04	.11	-.07	.04	.04
24	.02	.09	-.04	.00	.06	-.02
25	-.14	.03	-.01	-.05	.02	-.14
26	.08	.00	.17	.14	.05	.19
27	.11	.08	-.09	.14	.05	.16
28	.15	.02	.07	.04	.13	.10
29	.01	.05	-.16	.13	.07	.01
30	.05	.08	-.09	.09	.09	.08
31	.03	.13	-.07	.06	.34	.15
42	.18	.22	.06	-.05	.10	.17
Scale	Median correlation for					
Scale A	.02	.08	.10	.09	.10	.04
Scale B	.06	.06	-.02	.05	.06	.09
Both Scales	.05	.05	.06	.06	.07	.08

* Correlation coefficients which are underlined are significant at the .05 level. For Scale A, N equals 61. For Scale B, N equals 128.

TABLE 4
 MEDIAN CORRELATIONS BETWEEN PROJECTIVE TECHNIQUE RATINGS AND FINAL*
 POOLED RATINGS FOR EACH OF SIX PROJECTIVE RATERS

Scale	Rorschach Median correlation for												
	W	n	O	n	L	n	F	n	T	n	G	n	All
Scale A	.26	11	.10	11	.05	12	.16	10	.00	5	.43	9	.12
Scale B	.30	18	.21	23	.26	22	.44	22	.48	20	.20	20	.36
Both Scales	.30		.21		.22		.44		.48		.30		.19

Scale	TAT Median correlation for												
	W	n	O	n	L	n	F	n	T	n	G	n	All
Scale A	.71	7	.33	8	.37	9	.30	15	—	3	.30	14	.33
Scale B	.26	22	.16	22	.22	14	.44	28	.46	4	.23	25	.28
Both Scales	.28		.21		.26		.40		—		.28		.30

Scale	SC Median correlation for												
	W	n	O	n	L	n	F	n	T	n	G	n	All
Scale A	-.19	4	.54	8	.18	14	.82	4	.52	13	—	2	.52
Scale B	.04	16	.36	13	.32	37	.60	4	.45	26	.46	6	.36
Both Scales	.04		.47		.30		.74		.47		—		.37

*Correlation coefficients which are underlined are significant at the .05 level.

ing the same thing. Although the proportion of positive correlations is large, only 9 correlations out of 120 are significant at the .05 level, where 6 correlations are expected to reach this level by chance.

B. THE PROJECTIVE RATERS

1. Relative validities of ratings for individual raters

Beyond the techniques per se we are interested in the relative validities of the individual raters by technique and by

variable. In Table 4 may be found the median validity coefficients for all traits in Scale A and Scale B respectively, and for both Scales together. These are the median correlations for each of the six raters for ratings based on the Rorschach, TAT, and SC, respectively. Table 5 contains the same information for the Bender-Gestalt.

The results obtained (of which only the medians are presented here) suggest that two of the six Rorschach analysts

TABLE 5
 MEDIAN CORRELATIONS BETWEEN BENDER-GESTALT RATINGS AND FINAL POOLED
 RATINGS FOR EACH OF FOUR BENDER-GESTALT RATERS

Scale	Rater								
	R	n	Q	n	M	n	P	n	All
Scale A	.23	19	.00	14	.27	15	.30	13	.25
Scale B	.14	32	.10	29	.22	33	.12	33	.15
Both Scales	.14		.10		.26		.20		.16

were able to make ratings which correlated with the criterion ratings beyond chance expectancy at the .05 level. For these two raters, F and T, the median coefficients for Scale B are, respectively, .44 and .48. These are the only median coefficients which are statistically significant for the six raters on both scales. The number of statistically significant individual (not median) correlations is, considering both scales, greater than would be expected in chance occurrence for any rater, since at the .05 level only one in twenty coefficients is expected to reach this level by chance. The median value, however, is regarded as a better representation of the validity of a rater, in that it allows for the fact of intercorrelation between the traits within each Scale.

While the evaluation of a correlation coefficient in terms of what may be called its social significance is quite subjective, there is little reason for enthusiasm toward the efficiency with which these six Rorschach analysts were able to predict the criterion ratings. Although the validity coefficients for Rorschach raters may seem somewhat discouraging, it is emphasized that the ratings of Rorschach projectivists F and T appear to be reasonably valid. This finding is congruent with the situation which exists in the area of projective testing in which, to a very considerable extent, adeptness in the use of projective methods is acquired through instruction by masters. In the practical situation of assessment, however, four of the six Rorschach raters, initially selected on the basis of clinical competence, did not make as valid ratings as the other two. The high proportion of positive correlations suggests that, under assessment conditions, the Rorschach method provides a basis for some

validity of rating.

These findings and observations also apply, in general, to the data obtained for raters using the TAT, SC, and BG.

C. RELATIVE CONTRIBUTIONS OF RATERS, TECHNIQUES, AND TRAITS TO VALIDITIES OF RATINGS

In order to determine the relative contributions of each of the three sources of variance investigated in this study to the validities of ratings, the following analysis was carried out. The validity coefficients were regarded as scores, and, after being transformed into their respective z functions, were treated by the method of analysis of variance. Analyses were done separately for Scales A and B, and additional analyses for the Bender-Gestalt.

For Scale A traits the results indicated that significant differences in validity may be attributed both to techniques and to raters (.05 level) and that the interaction of techniques and raters makes for differences in validity (.01 level). That is to say, that for Scale A traits, the traits themselves do not contribute significantly to differences in validity. However, the different techniques make for differences in the validity of ratings and the different raters make for differences in validity of ratings. We find differences in validity dependent upon who rates, and upon what technique is being used, and differences in validity for the combinations of raters and techniques.

A similar analysis of Scale B traits showed significant differences in the validity of ratings for technique-rater combinations. Further analysis, however, showed this difference to be largely the contribution of rater differences in validity of rating. As with the Scale A traits, the traits themselves are not a source of differences in validity of ratings. While

the techniques did contribute to differences in rating validity for Scale A traits, they are not sources of difference for the Scale B traits. For the presumably covert aspects of personality it appears that differences in validity depend chiefly upon the rater.

Separate analyses of variance for each of the two Scales for the Bender-Gestalt showed that neither variance due to raters nor variance due to traits was significantly greater than variance due to chance. Since only the one projective method was involved in these analyses, there was no technique variance.

It seemed reasonable to expect that some personality traits would be rela-

tively more validly rated than others by one or all of the projective techniques or by one or all of the projective raters. The absence of statistically significant differences in validities of ratings attributable to differences among traits is necessarily qualified in that to the extent to which there is correlation among traits, it is more difficult to find existing differences. This is not a particularly cogent objection in the case of the Scale A variables for which the intercorrelations were low. The Scale B traits do have, in some instance, sizable correlation between them, resulting in an underestimation of the significance of the between-variable variance.

CHAPTER IV

OTHER FACTORS RELATED TO VALIDITY

IN THIS chapter are presented data with respect to factors, other than those originally considered in the scope of this study, which may have had an effect upon the validity of ratings based on projective techniques. Consideration will be given to the possibility of bias in rating; to differences in dispersion (confidence) of ratings; to some additional factors concerning raters; and to the possible effects of unreliability in the predictor and criterion ratings.

A. BIAS IN RATING

One factor which may be related to the validity of ratings based on projective techniques is the frame of reference in which ratings were made. By frame of reference is meant an attitude of optimism represented by a tendency to make ratings in the direction of the "laudatory" or socially desirable end of the rating scale. It is possible that if ratings on a given technique tend, consistently, to be either more or less "laudatory" than the Final Pooled Ratings, that this attitude might be related to the validity of such ratings. Similarly, it is possible that such differences in attitude might be related to differences between the relative validities of ratings among the techniques. In order to determine whether such "frame of reference" differences did obtain, the following was done. First, in order to compare the ratings by each of the techniques with the FinP ratings, the significance of the difference between the mean ratings was calculated. This was done by the usual method of obtaining the quotient of the difference between means divided by the standard error of the difference. In each instance

the FinP means were subtracted from the technique means and the arithmetic sign recorded in order that the direction of difference be known. This procedure was followed for the four techniques for the fifteen traits studies. A comparison of the FinP means with those of the Projective Integrator was necessarily limited to the ten Scale B traits. Second, the mean ratings by each of the techniques on the fifteen traits were compared with the mean ratings by each of the other projective techniques, again by calculating the significance of each of the differences.

The results of the comparison of mean ratings for techniques and FinP ratings are presented in Table 6 in the form of an abbreviated table of signs. Either a plus or a minus sign is entered in the table where a statistically significant (.05 level) difference between means occurs. The plus sign is used to indicate that the technique mean was in the "laudatory" direction, while the minus sign indicates that the FinP mean was in the "laudatory" direction. An examination of Table 6 reveals that the number of statistically significant differences found exceeds the number which might have been expected if only chance were involved but these differences are not consistently in the same direction. In the case of the Projective Integrator the obtained significant differences are always in the direction of being less "laudatory," and for the Rorschach one may note a trend in the same direction.¹ These data may be of interest insofar as

¹ Soskin (5) analyzed ratings based only on student behavior in a series of standard situations and found that Situationists "see the sample of subjects are being characterized predominately by the condemnatory pole. . . ."

TABLE 6
TABLE OF SIGNS SHOWING STATISTICALLY SIGNIFICANT DIFFERENCES BETWEEN
MEAN TECHNIQUE RATINGS AND MEAN FINAL POOLED RATINGS*

Variable	Ror-FinP	TAT-FinP	Sc-FinP	BG-FinP	PI-FinP
4					#
17	+		+		#
18					#
21					#
22		+			#
23					-
24	-				-
25				+	-
26			+		
27			-		
28					
29	-		-		-
30	-				-
31		+	+		
42					-

* In this table the "+" sign indicates that there is a statistically significant difference (.05 level) with the technique mean in the "laudatory" direction. The reverse is true for the "-" sign.

The Projective Integrator did not rate Scale A variables.

they characterize the PI and the Rorschach as tending to look on the dark side of personality, but probably have little, if any, relation to validity. When the differences shown in Table 6 were compared with the validity coefficients which are statistically significant for the techniques and for the Projective Integrator, it was apparent that statistically

significant validity coefficients were obtained at least as frequently when there was not a difference between means as when there was a mean difference. It may also be noted that statistically significant validity coefficients are not consistently associated with differences either toward or away from the "laudatory" end of the rating scale.

TABLE 7
TABLE OF STATISTICALLY SIGNIFICANT DIFFERENCES BETWEEN MEAN
RATINGS BASED ON EACH OF THE PROJECTIVE TECHNIQUES*

Variable	Ror-TAT	Ror-SC	Ror-BG	TAT-SC	TAT-BG	SC-BG
4						+
17						
18						
21						
22	-	-				
23			-	+		-
24			-			
25						
26		-	-	-		+
27		+		+		-
28				+		
29				+		-
30	-		-			
31						
42						-

* In this table the "+" sign indicates that there is a statistically significant difference (.05 level) with the first of the two techniques at the head of the column in the "laudatory" direction. The reverse is true for the "-" sign.

In Table 7 are presented the signs of statistically significant differences between the mean ratings among the techniques. In this table the minus sign indicates that the first of the two technique abbreviations at the head of the column has a mean rating in a less "laudatory" direction. Thus, the minus sign for Trait #22 in the first column signifies that

significant validity coefficients than the Bender-Gestalt without there being any differences between the means for these two techniques. Again, one might characterize the Rorschach raters as tending to view people darkly, while the Bender-Gestalt raters tend in the direction of rating personality in a relatively benign fashion.

TABLE 8

TABLE OF STATISTICALLY SIGNIFICANT DIFFERENCES BETWEEN STANDARD DEVIATIONS OF TECHNIQUE RATINGS AND FINAL POOLED RATINGS*

Variable	Ror-FinP	TAT-FinP	SC-FinP	BG-FinP	PI-FinP
4					#
17					#
18	+				#
21					#
22					#
23				+	
24	+				
25	+	+	+	+	
26				+	
27					
28		-			
29		+			
30					
31				+	
42			-		

* In this table the "+" sign indicates that there is a statistically significant difference (.05 level) with the technique ratings having the larger standard deviation. The reverse is true for the "-" sign.

The Projective Integrator did not rate Scale A variables.

there is a significant difference between the mean ratings of the Rorschach and TAT on this trait, and that the difference is in the direction of the Rorschach being less "laudatory." The apparent tendency for ratings based on the Bender-Gestalt to be more often in the "laudatory" direction than ratings based on the Rorschach or Sentence Completion may suggest a relationship to differences in validity, since both the Rorschach and Sentence Completion yielded more and higher statistically significant validity coefficients than the Bender-Gestalt. This argument is weakened, however, by the observation that the TAT too yielded more and higher statistically

B. DISPERSION (CONFIDENCE) OF RATINGS

Another factor which needs to be evaluated, since it might be related to validity of ratings, is the spread of scores or ratings. When there is more spread, the likelihood of obtaining higher correlation coefficients is increased. If one is willing to assume that where a rater lacks confidence in his rating he will tend to rate closer to the mean than he will when he does feel confident, then we can use a measure of spread as an index of confidence in evaluating confidence in rating as a factor in validity.

Table 8 is a table of signs showing where significant differences in spread (standard deviation) occur. A plus sign

TABLE 9
TABLE OF STATISTICALLY SIGNIFICANT DIFFERENCES BETWEEN STANDARD DEVIATIONS
OF RATINGS ON EACH OF THE PROJECTIVE TECHNIQUES*

Variable	Ror-TAT	Ror-SC	Ror-BG	TAT-SC	TAT-BG	SC-BG
4						
17						
18		+				
21						
22						
23					-	-
24		+				
25						
26						
27			-			
28					-	
29						-
30						
31		+		+	-	-
42	+	+	+			

* In this table the "+" sign indicates that there is a statistically significant difference (.05 level) with the first of the two techniques at the head of the column having the larger standard deviation. The reverse is true for the "-" sign.

indicates that the spread for the technique ratings is significantly greater than that for the FinP ratings, and vice versa for the minus sign. These data suggest that dispersion of ratings has little or no relation to validity. The ratings based on the Bender-Gestalt, which are the least valid of the four sets of ratings, most often show significantly greater spread than the FinP ratings whereas the Projective Integration ratings, which are more valid for more traits than the ratings based on the Bender-Gestalt, are in no instance significantly different in spread from the FinP ratings. It is interesting to note that ratings based on each of the four projective techniques (but not the Projective Integration) have significantly larger standard deviations on Trait #25, "Characteristic Intensity of Inner Emotional Tension," than the FinP. If this represents a feeling of confidence, it appears not to be justified consistently since the TAT and BG ratings on this trait did not correlate significantly with the criterion ratings.

A comparison of the techniques with

each other in terms of willingness to spread ratings was carried out. The results are presented in Table 9. The data suggest that ratings based on the Rorschach tend to be made with more spread than those based on the Sentence Completion and TAT. Similarly, the Bender-Gestalt tends to inspire dispersion of ratings at least equal to the Rorschach. This may be a function of the fact that both the Ror and BG were administered by the rater; i.e., face to face contact may be related to dispersion of ratings (confidence). This would be consistent with the staff's opinion that the interview contributed most to their understanding of the case.

C. ADDITIONAL FACTORS CONCERNING RATERS

It has been shown, (Chapter III, B, 2), that different raters contribute significantly to differences in validities of ratings. One possible explanation of this finding, that "good" raters may have been assigned to one technique and "poor" raters to another, is not too ten-

able since all the projective raters used the same techniques (except for the Bender-Gestalt) and, in addition, cases were randomly assigned to raters. On the other hand the number of cases seen by each projective rater varied and the obtained difference in validities for raters may have been related to difference in size of samples. As a way of evaluating whether there are differences in validity associated with the numbers of cases seen by "good" and "poor" raters, the rank-order positions of raters was used as an index of ability to rate. If we consider that those raters in the upper half of the ranking are "good" raters and those in the lower half are "poor" raters, then for Scale B, we find the following. For the Rorschach the "good" raters W, F, and T, saw 65 of a total of 125 cases; for the TAT the "good" raters, W, F, and T, saw 54 of the 115 cases; for the SC the "good" raters, F, T, and G, saw 36 of 102 cases. Also, the size of the median validity coefficient for the SC compares favorably with those of the Rorschach and TAT. We may conclude that the size of sample for different raters is not significantly related to validity of rating.

Another possibility to be considered is the effect of a given projective technique on a rater. There may be something in the nature of a given technique which effects a change in the clinician who uses it. The data in Table 6 suggest that the Rorschach is more often associated with a tendency to see people in a less benign light than are the other techniques. When the Rorschach rater in his role of Projective Integrator makes a new set of ratings, incorporating the results of the other projective techniques with his own Rorschach results, this tendency to see people in a less benign light is increased.

D. FURTHER EVIDENCE ON INTER-JUDGE AND INTER-TECHNIQUE AGREEMENT

As was reported in Chapter III, A, 1, this study found that unreliability of criterion ratings has very little effect on the validity of ratings based on projective techniques. There is no way of knowing, from the present data, to what extent the low validities found are a function of the unreliability of raters. The question may be asked in two ways: to what degree do judges agree when they use the same technique, to what degree does a judge agree with himself using different techniques? In order to answer these questions another study was carried out and some preliminary results were available at the time of writing. In this new study four projectivists made ratings based on each of four projective technique records from each of a total of 20 subjects. The number of subjects and raters was limited by practical considerations. On each of the 16 days of the experiment a rater made ratings based on five projective-technique records from 5 different subjects. For example, on the first day Rater A would have worked with the Rorschach records of Subjects 1 and 16, the TAT of Subject 10, the SC of Subject 12, and the B-G of Subject 4. This same "block" of records would be given to Rater B on the eighth day, to Rater C on the fifth day, and to Rater D on the sixteenth day. In this way each of the 16 blocks of five projective-technique records would be worked with by each of the four raters, so that at the end of the 16 days each rater would have rated all the projective records on all 20 subjects. Tables of random numbers were used to assign the projective technique records to the 16 "blocks," after which the blocks were treated as units and distributed by random numbers to the raters.

TABLE 10

MEDIAN INTERCORRELATION COEFFICIENTS FOR EACH OF FOUR RATERS WHEN RATINGS BASED ON FOUR TECHNIQUES ARE INTERCORRELATED*

Variable	Rater			
	A	B	C	D
4	.32	.12	.19	.06
18	.08	.31	.23	.17
23	-.06	.20	-.03	.02
25	.04	-.15	.06	.18
31	.02	.32	.51	.20
42	.05	.26	.16	.14

* Correlation coefficients which are underlined are significant at the .05 level.

The ratings were then examined for consistency of rating for a rater over the four techniques, and for the consistency of ratings on a given technique by the four raters, in both cases for 20 subjects.

In Table 10 are shown the results for the consistency of the individual raters, i.e., the intra-rater consistency over all four techniques. The data are presented in the form of median values for the 6 intercorrelations of the four techniques. The single correlation which appears to be significant (Rater C, Trait #31) may be meaningless in that at the .05 level 1 correlation in 20 may be expected to reach this level of significance.

In Table 11 are shown the inter-rater

TABLE 11

MEDIAN INTERCORRELATION COEFFICIENTS FOR EACH OF FOUR TECHNIQUES WHEN RATINGS BY EACH RATER ON THAT TECHNIQUE ARE INTERCORRELATED*

Variable	Technique			
	Ror	TAT	SC	BG
4	.28	.49	.34	.18
18	.52	.22	.78	.44
23	.35	.20	.36	.20
25	.22	.40	.17	.20
31	.58	.16	.51	.34
42	.56	.34	.34	.48

* Correlation coefficients which are underlined are significant at the .05 level.

(intra-technique) agreements. These again are median correlations for the 6 intercorrelations of the ratings of four raters for the same technique. It appears, from the data in these two tables, that a projective technique tends to generate more consistency in rating behavior than the use of a single projective rater. Stated another way, it would appear that different raters will tend to rate a subject in the same way if they use the same projective technique. A given rater will tend to rate the same subject in a nonconsistent manner when he uses different techniques.

CHAPTER V

SUMMARY AND CONCLUSIONS

A. OBJECTIVES

THE principal objectives of this study were:

1. To determine the relative validities of ratings based on four projective techniques when these were used to describe personality and to predict "over-all suitability" as a clinical psychologist;
2. To determine whether there were differences in the relative validities of ratings based on projective techniques which were ascribable to the rater;
3. To determine whether there were differences in the relative validities of ratings based on projective techniques which were ascribable to the technique;
4. To determine whether there were differences in the relative validities of ratings based on projective techniques which were attributable to the personality trait being rated.

Other related problems were also investigated.

B. METHODS AND PROCEDURES

During an intensive assessment program, a total of 128 male first-year graduate students in clinical psychology were rated by ten staff members on 15 personality variables after analysis of protocols from each of four projective techniques, and again after all the projective material (except the actual ratings) from the four techniques had been examined by a "projective integrator." These ratings were correlated with another set of ratings which were arrived at in a "pooling" conference of three staff members who had studied each subject intensively for a week. Since staff members had all the information it was possible to obtain

about a subject during an assessment week, the ratings which the three staff members agreed upon in their final conference were taken as the best-rated descriptions of a subject's personality and used as criterion measures. These criterion measures were "contaminated" in that they included the protocols and ratings of the predictors, the effect of which would be spuriously to raise the correlations between ratings based on projective techniques and criterion measures. Under these conditions the correlations between ratings based on projective techniques and the final staff ratings were regarded as validity indices. By comparing the correlation coefficients which were obtained in this manner, the objectives of this study could be achieved.

C. SUMMARY OF FINDINGS

1. *The projective techniques.*

- a. Ratings based on the projective techniques correlated significantly with the criterion measures more frequently than was expected by chance. These correlations were preponderantly positive, but were, by usual standards, low.
- b. Correcting the criterion measures for attenuation raised the median correlations between ratings based on projective techniques and the criterion measures only slightly. The greatest increase was from .31 to .35.
- c. Ratings based on the Rorschach and the Bender-Gestalt (made by the raters who had administered these techniques) tended to be made with greater confidence, i.e.,

showed greater spread, than ratings based on the Thematic Apperception Test and the Sentence Completion (which were rated "blind").

- d. Ratings based on the Rorschach tended to be made away from the "laudatory" end of the ratings scale when compared with ratings based on the other projective techniques and with the criterion ratings.
 - e. There appeared to be little or no relation between validity of ratings and bias of ratings toward either the "laudatory" or "derogatory" ends of the rating scale.
 - f. There appeared to be little or no relation between validity of ratings and spread of ratings.
 - g. Preliminary findings suggested that there was more correlation between ratings made by different raters using the same projective technique, than between ratings made by a single rater using different projective techniques.
2. *The projective raters.*
- a. Differences in the validity of ratings based on projective techniques were found which could be attributed to the individuals making the ratings.
 - b. Significant differences in the validity of ratings obtained not only between raters, but for the interaction between rater and technique, for both scales and for the Rorschach, the Thematic Apperception Test, and the Sentence Completion; interaction could not

be tested for the Bender-Gestalt.

- c. Although some raters were "better" than others, in no instance did the median validity coefficients of statistical significance exceed $r = .48$.

3. *The Traits.*

No significant differences were found in the validity of ratings which were attributable to the differences in personality traits rated.

D. CONCLUSIONS

The conclusions which are drawn from the findings of this study are:

1. Projective techniques, used in the assessment of personality characteristics, measure very little in common.
2. There are significant individual differences in the ability to make valid ratings of personality traits from projective techniques, which appear to be independent of the technique used.
3. The dispersion of ratings made by clinicians on personality traits on the basis of projective techniques does not appear to be related to the validity of ratings.
4. The assessment of the degree to which a subject possesses socially desirable personality traits is, in part, a function of the projective technique which is used in making the assessment.
5. The value of projective techniques as instruments for the assessment of specified personality traits in a group of normal superior adults is apparently limited by low validities.

BIBLIOGRAPHY

1. CATTELL, R. B. *Description and measurement of personality*. Yonkers-on-Hudson: World Book Co., 1946.
2. FISKE, D. W. Consistency of factorial structures of personality ratings from different sources. *J. abnorm. soc. Psychol.* 1949, **44**, 329-344.
3. KELLY, E. L. Research on the selection of clinical psychologists. *J. clin. Psychol.* 1947, **3**, 39-42.
4. KELLY, E. L. & FISKE, D. W. The prediction of success in the VA training program in clinical psychology. *Amer. Psychologist*, 1950, **5**, 395-406.
5. SAMUELS, H. An analysis of some factors affecting ratings of personality traits based on projective techniques. Ph.D. Thesis, Univer. Mich., 1956.
6. SOSKIN, W. F. A study of personality ratings based upon brief observations of behavior in standard situations. *Microfilm Abstr.* (Ph.D. Thesis) University Microfilms, Ann Arbor, Mich. Publ. No. 1208.
7. TUPES, E. C. An evaluation of personality trait ratings obtained by unstructured assessment interviews. *Psychol. Monogr.* 1950, **64**, No. 11 (Whole No. 317).